CS5500

Reinforcement Learning (RL)

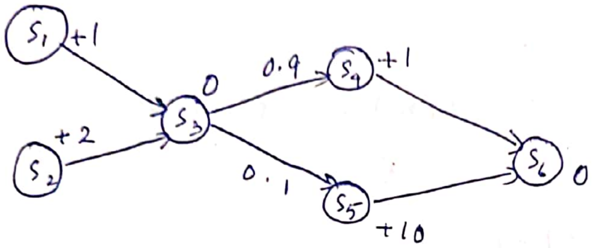
Assignment No. 2

Name- Raktim Gautam Goswami

Roll number - EE17BTECH11051

Q1)

Sol^n-



a) $V(S_6) = 0$

$V(S_4) = +1 + 0 = 1$

$V(S_5) = 10 + 0 = 10$

$V(S_3) = 0.9 \times 1 + 0.1 \times 10 + 0 = 0.9 + 1 = 1.9$

$V(S_2) = 2 + V(S_3) = 2 + 1.9 = 3.9$

$V(S_1) = 1 + V(S_3) = 1 + 1.9 = 2.9$

For (b), (c), (d) & (e), the trajectories along with the rewards are written as follows -

(1) $S_1(1) \rightarrow S_3(0) \rightarrow S_4(1) \rightarrow S_6(0)$

(2) $S_1(1) \rightarrow S_3(0) \rightarrow S_5(10) \rightarrow S_6(0)$

(3) $S_1(1) \rightarrow S_3(0) \rightarrow S_4(1) \rightarrow S_6(0)$

(4) $S_1(1) \rightarrow S_3(0) \rightarrow S_4(1) \rightarrow S_6(0)$

(5) $S_2(2) \rightarrow S_3(0) \rightarrow S_5(10) \rightarrow S_6(0)$

b) $V(S_1) = \frac{1}{4}(2 + 11 + 2 + 2) = \frac{17}{4} = 4.25$

$V(S_2) = \frac{12}{1} = 12$

c) Let us initialize $V(s_1) = V(s_2) = V(s_3) = V(s_4) = V(s_5) = V(s_6) = 0$

$\alpha_t = \frac{1}{t}$

After first episode,

$V(s_6) = 0$

~~$V(s_5) = 0 + \frac{1}{1}(10+0-0) = 10$~~

$V(s_4) = 0 + \frac{1}{1}(1+0-0) = 1$

$V(s_3) = 0 + \frac{1}{1}(0+1-0) = 1$

$V(s_1) = 0 + \frac{1}{1}(1+1-0) = 2$

After second episode,

$V(s_6) = 0$

$V(s_5) = 0 + \frac{1}{2}(10+0-0) = 5$

$V(s_3) = 1 + \frac{1}{2}(0+5-1) = 1+2 = 3$

$V(s_1) = 2 + \frac{1}{2}(1+3-2) = 3$

After third episode,

$V(s_6) = 0$

$V(s_4) = 1 + \frac{1}{3}(1+0-1) = 1$

$V(s_3) = 3 + \frac{1}{3}(0+1-3) = 3 - \frac{2}{3} = \frac{1}{3}$

$V(s_1) = 3 + \frac{1}{3}(1 + \frac{1}{3} - 3) = 3 - \frac{5}{9} = \frac{22}{9}$

After fourth episode,

$V(s_6) = 0$

$V(s_4) = 1 + \frac{1}{4}(1+0-1) = 1$

$V(s_3) = \frac{1}{3} + \frac{1}{4}(0+1-\frac{1}{3}) = \frac{1}{2}$

$V(s_1) = \frac{22}{9} + \frac{1}{4}(1 + \frac{1}{2} - \frac{22}{9}) = \frac{53}{24}$

After fifth episode,

$V(s_6) = 0$

$V(s_5) = 5 + \frac{1}{5}(10+0-5) = 6$

$V(s_3) = \frac{1}{2} + \frac{1}{5}(0+6-\frac{1}{2}) = \frac{8}{5}$

$V(s_2) = 0 + \frac{1}{5}(2 + \frac{8}{5} - 0) = \frac{18}{25}$

$\therefore V(s_1) = \frac{53}{24} \simeq 2.208$

$V(s_2) = 0.72$

d) From the given samples,

$P(s' = s_4 \mid s = s_3) = \frac{3}{5} = 0.6$

$P(s' = s_5 \mid s = s_3) = \frac{2}{5} = 0.4$

$V(s_6) = 0 \quad ; \quad V(s_4) = 1 \quad ; \quad V(s_5) = 10$

$V(s_3) = \frac{3}{5} \times 1 + \frac{2}{5} \times 10 = \frac{23}{5}$

$\therefore V(s_2) = 2 + \frac{23}{5} = \frac{33}{5} = 6.6$

c) The estimate of $V(s_2)$ by TD(0) method is the closest to the true value. MC estimate is far from the true value. This is because → there is bootstrapping in TD but not in MC

→ The number of episodes is very few for MC to work efficiently

Q2)

sol^n - (1) $\alpha_t = \frac{1}{t}$

$$\sum_{t=0}^{\infty} \alpha_t = \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} + \cdots$$

$$= 1 + \left(\frac{1}{2}\right) + \left(\frac{1}{3} + \frac{1}{4}\right) + \left(\frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8}\right) + \cdots$$

$$= 1 + 0.5 + 0.58 + 0.63 + \cdots$$

Each of the terms above are greater than 0.5 and there are infinite such terms.

$$\therefore \sum_{t=0}^{\infty} \alpha_t \to \infty$$

$$\sum_{t=1}^{\infty} \alpha_t^2 = \sum_{t=1}^{\infty} \frac{1}{t^2}$$

$$\frac{1}{t^2} < \frac{1}{t(t-1)} = \frac{1}{t-1} - \frac{1}{t}$$

$$\therefore \sum_{t=1}^{\infty} \frac{1}{t^2} < \sum_{t=2}^{\infty} \left(\frac{1}{t-1} - \frac{1}{t}\right)$$

$$= \lim_{n\to\infty} \sum_{t=0}^{n} \left(\frac{1}{t-1} - \frac{1}{t}\right)$$

$$= \lim_{n\to\infty} \left(\frac{1}{1} - \cancel{\frac{1}{2}} + \cancel{\frac{1}{2}} - \cancel{\frac{1}{3}} + \cancel{\frac{1}{3}} - \cancel{\frac{1}{4}} + \cancel{\frac{1}{4}} \cdots - \frac{1}{n}\right)$$

$$= 1$$ ∴ this will result in convergency

(2) $\alpha_t = \frac{1}{t^2}$

In previous part we have proved that

$$\sum_{t=1}^{\infty} \frac{1}{t^2} < 1$$

$$\Rightarrow \sum_{t=1}^{\infty} \alpha_t < 1$$

$\therefore$ $\alpha_t = \frac{1}{t^2}$ will not result in convergence

(2) $\alpha_t = \frac{1}{t^{1/2}}$

$$\sum_{t=1}^{\infty} \alpha_t^2 = \sum_{t=1}^{\infty} \frac{1}{t} \rightarrow \infty$$

$\therefore$ $\alpha_t = \frac{1}{t^{1/2}}$ will not result in convergence

(3) $\alpha_t = \frac{1}{t^{2/3}}$

~~Let~~ $\cancel{S_{2k+1} = \sum_{n=1}^{2k+1} \frac{1}{t^{2/3}}}$

$$\sum_{t=1}^{\infty} \frac{1}{t^{2/3}} = \frac{1}{1} + \frac{1}{2^{2/3}} + \frac{1}{3^{2/3}} + \frac{1}{4^{2/3}} + \frac{1}{5^{2/3}} + \cdots$$

$$> \frac{1}{1} + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \frac{1}{5} + \cdots$$

$$= \infty$$

$$\therefore \sum_{t=1}^{\infty} \frac{1}{t^{2/3}} \rightarrow \infty$$

Now,

$$\sum_{t=1}^{\infty} \alpha_t^2 = \sum_{t=1}^{\infty} \frac{1}{t^{4/3}}$$

We know that this will converge if

$\sum_{k=0}^{\infty} \frac{2^k}{2^{4/3 k}}$ converges.

$$\therefore \sum_{k=0}^{\infty} 2^{-k/3} = \frac{1}{1-2^{-1/3}} < \infty$$

$\therefore$ $\alpha_t = \frac{1}{t^{2/3}}$ will result in convergence.

Now, for any integer p, $p > 0$

$$\alpha_t = \frac{1}{t^p}$$

We can show its convergence / divergence using integral test.

$$\therefore \int_1^{\infty} \frac{1}{t^p}\,dt = \lim_{D\to\infty} \left.\frac{x^{1-p}}{1-p}\right|_1^D = \lim_{D\to\infty} \frac{D^{1-p}}{1-p} - \frac{1}{1-p}$$

This converges when $1-p < 0$

$$\Rightarrow p > 1$$

$\therefore$ it diverges when $p < 1$

$\sum_{t=1}^{\infty} \frac{1}{t^{2p}}$ converges when $1-2p < 0$

$$\Rightarrow p > \frac{1}{2}$$

$\therefore$

$\frac{1}{2} < p < 1$ would result in convergence.

Q3)

Sol^n:

$$Q_\pi(s, \pi'(s)) = \sum_{a\in A} \pi'(a|s) Q_\pi(s,a)$$

$$= \epsilon/m \sum_{a\in A} Q_\pi(s,a) + (1-\epsilon) \max_{a\in A} Q_\pi(s,a)$$

$$\geq \epsilon/m \sum_{a\in A} Q_\pi(s,a) + (1-\epsilon) \sum_{a\in A} \frac{\pi(a|s) - \epsilon/m}{1-\epsilon} Q_\pi(s,a)$$

$$= \sum_{a\in A} \pi(a|s) Q_\pi(s,a)$$

$$= V^{\pi}(s)$$

$$\therefore V^{\pi'}(s) \geq V^{\pi}(s)$$

Q4)

Sol^n:

$$G_t^{\lambda} = (1-\lambda)\left[\lambda^0 G_t^{(1)} + \lambda^1 G_t^{(2)} + \lambda^2 G_t^{(3)} + \dots + \lambda^{n-1} G_t^{(n)}\right]$$

Let the k^{th} weight be such that

$$\frac{\lambda^{k-1}}{\lambda^0} < \frac{1}{2}$$

$$\Rightarrow (k-1)\ln(\lambda) < \ln(0.5)$$

$\Rightarrow \quad K < \dfrac{-0.693}{\ln(\lambda)}$

$\Rightarrow \quad n(\lambda) = \dfrac{-0.693}{\ln(\lambda)}$

$\therefore$ for $\quad n(\lambda) = 3$

$\Rightarrow \quad -\dfrac{0.693}{\ln(\lambda)} = 3$

$\Rightarrow \quad \lambda = e^{-0.23105}$

$= 0.7937$

Q5)

Solⁿ- On ~~[illegible]~~ writing a program for the given MDP and running it, we get the following policy on convergence -

if $s = s_1$, $a = a_3$

$s = s_2$, $a = a_3$

$s = s_3$, $a = a_3$

~~∴ the Q-learner~~ The given trajectory is $(s_1, a_1, 1, s_1, a_2, 2, s_2)$

This is possible because the agent follows ϵ-greedy approach. Both the actions in this trajectory are random.